CME 323: Distributed Algorithms and Optimization

Instructor: Reza Zadeh (rezab@stanford.edu)

HW#4 - Due Thursday, June 6

1. **Shallow Graphs** For an undirected graph $G = (V, E)$ with n vertices and m edges ($m \geq n$), we say that G is shallow if for every pair of vertices $u, v \in V$, there is a path from u to v of length at most 2 (i.e. using at most two edges).

   (a) Give an algorithm that can decide whether G is shallow in $O(n^{2.376})$ time.

   (b) Given an $n \times r$ matrix A and an $r \times n$ matrix B where $r \leq n$, show that we can multiply A and B in $O\left((n/r)^2 r^{2.376}\right)$ time. Hint: use the fact that we can multiply two $r \times r$ matrices in $O(r^{2.376})$ time.

   (c) Give an algorithm that can decide whether G is shallow in $O(m^{0.55} n^{1.45})$ time. Hint: consider length-2 paths that go from low-degree vertices and length-2 paths that go through high-degree vertices separately. Use result from part (b).

2. Write a Spark program to compute the Singular Value Decomposition of the following 10×3 matrix:

```
-0.5529181 -0.5465480 0.009519836
-0.5428579 -1.5623879 0.982464609
-1.3038629  0.5715549 0.499441144
 0.6564096  1.1806877 0.495705999
-1.2061171  1.3430651 0.153477135
 0.2938439 -1.7966043 0.914381381
-0.2578953  0.2596407 0.815623895
 0.9659582  2.3697927 0.320880634
-0.4038109  0.9846071 0.488856619
 0.6029003 -0.3202214 0.380347546
```

   Assume the matrix is tall and skinny, so the rows should be split up and inserted into an RDD. Each row can fit in memory on a single machine. Report all singular vectors and values and submit your Spark program.

3. Given a matrix M in row format as an RDD[Array[Double]] and a local vector x given as an Array[Double], give Spark code to compute the matrix vector multiply Mx.

4. In class we saw how to compute highly similar pairs of m-dimensional vectors x, y via sampling in the mappers, where the similarity was defined by cosine similarity: $\frac{x^T y}{|x|_2 |y|_2}$. Show how to modify the sampling scheme to work with overlap similarity, defined as

$$\text{overlap}(x, y) = \frac{x^T y}{\min(|x|_2^2, |y|_2^2)}$$

   (a) Prove shuffle size is still independent of m, the dimension of x and y.

   (b) Assuming combiners are used with B mapper machines, analyze the shuffle size.